International Journal of Forecasting 33 (2017) 314–324

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Using a rolling training approach to improve judgmental extrapolations elicited from forecasters with technical knowledge

Fotios Petropoulos [a,*], Paul Goodwin [b], Robert Fildes [c]

[a] Cardiff Business School, Cardiff University, UK
[b] School of Management, University of Bath, UK
[c] Lancaster Centre for Forecasting, Lancaster University, UK

ARTICLE INFO

Keywords:
Judgmental forecasting
Unaided judgments
Rolling training
Feedback
Time series
Expert knowledge elicitation

ABSTRACT

There are several biases and inefficiencies that are commonly associated with the judgmental extrapolation of time series, even when the forecasters have technical knowledge about forecasting. This study examines the effectiveness of using a rolling training approach, based on feedback, to improve the accuracy of forecasts elicited from people with such knowledge. In an experiment, forecasters were asked to make multiple judgmental extrapolations for a set of time series from different time origins. For each series in turn, the participants were either unaided or provided with feedback. In the latter case, the true outcomes and performance feedback were provided following the submission of each set of forecasts. The objective was to provide a training scheme that would enable forecasters to understand the underlying pattern of the data better by learning from their forecast errors directly. An analysis of the results indicated that this rolling training approach is an effective method for enhancing the judgmental extrapolations elicited from people with technical knowledge, especially when bias feedback is provided. As such, it could be a valuable element in the design of software systems that are intended to support expert knowledge elicitation (EKE) in forecasting.

1. Introduction

Surveys suggest that forecasts based either wholly or partly on expert management judgment play a major role in company decision making (e.g., Fildes & Goodwin, 2007). Sometimes these judgmental inputs take the form of adjustments to statistical forecasts, ostensibly to take into account special factors that were not considered by the statistical forecast (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009). However, in some circumstances, judgment may be the only process involved in producing the forecasts. At times, there is even a statistical forecast provided, but the expert chooses to ignore it (Franses, 2014). In some cases, judgment is used to extrapolate time series data to produce point forecasts, when no other information (except perhaps variable labels such as 'sales' or 'costs') is provided. This type of task has been the subject of much research over the last thirty years, and a number of biases associated with judgmental extrapolation have been identified. These include tendencies to overweight the most recent observation (e.g., O'Connor, Remus, & Griggs, 1993, to underestimate the growth and decay in

* Corresponding author.
E-mail address: PetropoulosF@cardiff.ac.uk (F. Petropoulos).

http://dx.doi.org/10.1016/j.ijforecast.2015.12.006

series (Lawrence, Goodwin, O'Connor, & Onkal, 2006), and to see systematic patterns in the noise associated with series (Eggleton, 1982; O'Connor et al., 1993).

Such biases can apply even when the forecaster has expertise, whether in the domain within which the forecasts are being made (e.g., Pollock & Wilkie, 1993) or in forecasting itself (Goodwin & Fildes, 1999). This suggests that, when experts are called upon to make judgmental extrapolations, the elicitation process may benefit from the inclusion of devices that are designed to mitigate these biases. Studies in the expert knowledge and elicitation (EKE) literature have examined a number of ways of designing elicitation methods so as to reduce the danger of biased judgments from experts, particularly in relation to the estimation of probabilities or probability distributions Aspinall, 2010; Bolger & Rowe, 2014; Goodwin & Wright, 2014, Morgan, 2014, Chapter 11). Our focus here is on improving EKE in time series extrapolation.

A variety of strategies have been explored in an attempt to mitigate biases in the elicitation of judgmental extrapolations (Goodwin & Wright, 1993). One promising strategy is to use performance feedback to train forecasters who already have technical expertise in order to improve the accuracy of their extrapolations (Lawrence et al., 2006). The use of feedback to enhance the quality of expert judgments has proved to be successful in other areas of EKE, such as weather forecasting (Murphy & Winkler, 1977), as well as in applications of the Delphi technique, where the feedback relates to the judgments of other experts (Rowe & Wright, 1999). In time series extrapolation, while some studies, such as that of Goodwin and Fildes (1999), have shown that feedback can lead to improvements in the accuracy of point forecasts, more research is needed to identify the most effective form of feedback for improving the accuracy. This is a particularly important topic in demand forecasting, where software provides the expert with information on past errors.

This paper reports on an experiment that was designed to examine the effectiveness of providing forecasters with rolling feedback on both the outcomes of the variable that they are attempting to predict and their forecasting performance. The objective is to provide a direct training scheme, thus enabling forecasters who already have technical knowledge to understand the underlying pattern of the data better by learning from their forecast errors directly, thus improving the accuracy of their judgments. Two types of performance feedback were compared: feedback on the bias associated with the forecasts submitted, and feedback on their accuracy. The paper is structured as follows. First, a review of the relevant literature is presented. Details of the experiment and the analysis and results follow. Finally, the practical implications of the findings are discussed, and suggestions are made for further work in this area.

2. Literature review

In judgmental forecasting, Sanders and Ritzman (1992) distinguish between expertise that is founded on contextual knowledge and that which is based on technical knowledge. Expertise relating to contextual knowledge comes from factors such as experience working in an industry or the possession of specific product knowledge. In contrast, expertise based on technical knowledge is present when a forecaster has a knowledge of formal forecasting procedures, including information on how to analyze data judgmentally.

Sanders and Ritzman compared the forecasting accuracies of: (i) managers who had contextual expertise but lacked technical expertise, (ii) forecasters who lacked contextual expertise but had technical expertise, and (iii) forecasters who lacked both contextual and technical expertise. They concluded that expertise based on technical knowledge had little value in improving the accuracy of judgmental forecasts relative to expertise based on contextual knowledge. However, many of the time series that they studied were highly volatile, and contextual factors, rather than time series components, accounted for much of their variation. The forecasters with technical expertise who took part in the study were not privy to these contextual factors.

A comparison of the forecasts of people in groups (ii) and (iii) enabled the authors to assess whether forecasters who were lacking in contextual expertise but educated in such technical aspects as the handling of outliers, the identification of trends and the avoidance of judgmental biases were able to achieve higher levels of accuracy than those who lacked such knowledge. The authors reported that there was little difference in accuracy, and therefore concluded that providing people with technical expertise had no value. However, a close inspection of their results reveals that this finding only holds for the five most volatile series in the study (those with a coefficient of variation exceeding 134%). If these series are excluded, forecasters with technical knowledge had lower average median absolute percentage errors (MdAPE) than those without this knowledge in 13 series out of 17 ($p = 0.025$ on a binomial test of the hypothesis that each group had an equal probability of achieving the lowest MdAPE on a given series). Although the mean reduction in average MdAPEs for the 17 series was only 1.8%, the results provide some evidence that, when series do not demonstrate extreme volatility, there may actually be advantages in eliciting forecasts from people who possess technical expertise. This also suggests that it may be possible for these judgments to be enhanced through further training.

In a review of Sanders and Ritzman's (1992) study, Collopy (1994) argues that people may not always be able to apply what they learn in a training process. He cites a report by Culotta (1992), who found that even students who do well in calculus courses cannot apply what they have learned. In Sanders and Ritzman's study, those who were counted as having technical knowledge had taken an elective course in forecasting, and may therefore have been subject to didactic learning, which is a relatively passive process. This is in contrast to experiential learning, which includes actively participating in the task for which one is being trained, reflecting on the experience, and learning from feedback (Moon, 2004). Thus, training of this type may be effective in obtaining improvements in accuracy for those with technical expertise.

In order for experiential training to be effective, it needs to address the specific challenges of the task

(Kremer, Moritz, & Siemsen, 2011). Goodwin and Wright (1993, 1994) argue that three components of a time series influence the degree of difficulty that is associated with the judgmental time series forecasting task, namely: (1) the complexity of the underlying signal, comprising factors such as its seasonality, cycles and trends, and autocorrelation; (2) the level of noise around the signal; and (3) the stability of the underlying signal.

When there are trends in series, studies have consistently found that judgmental forecasters tend to damp them when making extrapolations (Eggleton, 1982; Lawrence & Makridakis, 1989; O'Connor, Remus, & Griggs, 1997). This phenomenon appears to apply both to experts working in their specialist field and to participants in experimental studies (e.g., Wagenaar & Sagaria, 1975). This damping may occur either because forecasters anchor on the most recent observation and make insufficient adjustments from this (Bolger & Harvey, 1993), or because they are unable to handle non-linear change. However, damping may also be caused by forecasters bringing non-time series information, based on their knowledge or experience, to the task. For example, a forecaster's prior experience may have demonstrated that the sales growth for products tends to be damped. Similarly, in the case of a downward trend in a sales series, people may expect a trend reversal to occur as action is taken to correct the decline (O'Connor et al., 1997). Complex seasonal patterns or cyclical components have also been found to lead to inaccurate judgmental forecasts (Lawrence & O'Connor, 1993).

Several studies have suggested that judgmental forecasters often confuse the noise in the time series with the signal (Andreassen, 1988; Harvey, 1995; Lopes & Oden, 1987; Reimers & Harvey, 2011). For example, they often adjust statistical forecasts to take into account recent random movements in series which they perceive to be systematic changes that were not detected by the statistical forecast (Goodwin & Fildes, 1999). Conversely, when systematic changes in the signal do occur, forecasters may delay their responses, perceiving the changes to be noise (O'Connor et al., 1993). Also, they may pay too much attention to the most recent observation, which will contain a certain amount of noise (Bolger & Harvey, 1993; Lawrence & O'Connor, 1992). It seems reasonable to expect that noise could also impair the detection of underlying trends and seasonal patterns, though this was not the case in two studies where the series were presented graphically (Lawrence & Makridakis, 1989; Mosteller, Siegel, Trapido, & Youtz, 1981).

Learning through feedback could potentially mitigate these biases (Lawrence et al., 2006). As we indicated above, feedback is a key component of experiential learning, and has been shown to improve the accuracy of point forecasts (Goodwin & Fildes, 1999; Remus, O'Connor, & Griggs, 1996; Sanders, 1997; Welch, Bretschneider, & Rohrbaugh, 1998). However, there are a number of different types of feedback that may be particularly relevant to the task of time series forecasting (Balzer, Doherty, & O'Connor, 1989; Önkal & Muradoglu, 1995), and more research is needed to determine the type that is most effective and how it should be delivered.

The simplest form is outcome feedback, where the forecaster is told the outcome of the variable that they have been forecasting as it becomes available. This allows them to compare each forecast with outcome directly, which may help them to improve their forecasting accuracy over time. However, there is evidence that learning through outcome feedback can be slow (Klayman, 1988). One problem is that each outcome will contain an element of noise, and therefore highlighting this may exacerbate a forecaster's tendency to pay too much attention to the latest observation and to overreact to noise in the series. However, this may not be the case when outcomes are provided for a set of periods ($n > 1$), rather than just one period. In any case, outcome feedback is easy to provide, easy to understand, and not contaminated by older and possibly irrelevant observations (Goodwin, Onkal-Atay, Thomson, Pollock, & Macaulay, 2004). It is probably also something that forecasters would naturally expect to see, so it seems reasonable to supply it even if other forms of feedback are being provided as well.

Performance feedback provides forecasters with information on the quality of their forecasts, such as their accuracy or any bias. It usually takes the form of an average, reflecting the forecaster's performance over a number of periods. Determining the number of periods over which to average the performance poses a dilemma: too few, and the feedback may be based on too small a sample of forecasts to provide a reliable assessment of performance; too many, and the performance measure will not reflect recent improvements or deteriorations in performance adequately. The use of an exponentially weighted moving average of performances may help to solve the dilemma, but may be less transparent and understandable to the recipients of the feedback. Another option would be to simply supply a set of point errors for n recent periods without using any kind of average. This could potentially enable the forecaster to identify specific problematic periods that require attention (for example, seasonality peaks). Moreover, when using a rolling origin scheme, this strategy provides a way to check whether the point errors are decreasing over time.

We might expect the effectiveness of different types of performance feedback to vary. Feedback on biases can provide a direct message that one's forecasts are typically too high or too low, hence suggesting how they might be improved. This is likely to be beneficial for untrended series or series with monotonic trends. However, it may lead to an unwarranted confidence in one's current forecasting strategy when a series has an alternating or seasonal pattern, because biases in different periods will tend to cancel each other out if an average across the signed errors is used. In contrast, feedback on accuracy provides no such direct message, and its implications may be difficult to discern. In order for forecasters to learn from accuracy feedback, they would need to experiment with alternative approaches, not specified by the feedback, and then establish whether these have improved the accuracy. This requires forecasters to compare their levels of accuracy across different periods, which adds to their cognitive burden. Thus, it seems unlikely that accuracy feedback will be conducive to rapid learning. This may explain the ineffectiveness of performance feedback that was found in a study by Remus et al. (1996), which

consisted only of an accuracy measure (the mean absolute percentage error).

Other forms of feedback seem likely to be less relevant to practical judgmental time series forecasting contexts. Cognitive process feedback aims to provide forecasters with insights into their own forecasting strategies, causing them to reflect on the possible deficiencies of these strategies (O'Connor, Remus, & Lim, 2005). For example, a regression model may be used to attempt to capture their strategy, to allow the identification of the weights that are implicitly being attached to different items of available information, or cues. In time series forecasting, it will clearly take time to obtain sufficient information to enable these weights to be estimated reliably, which reduces the speed at which forecasters can learn. Also, identifying the cues that should be included in a model from the huge number of potential cues that are present in the time series forecasting task is problematic (e.g., typical cues might be the last observation, the mean of the last n observations, the last difference between observations, the range of the last n observations, and so on). In addition, many of these cues will be serially correlated, meaning that multicollinearity is likely to reduce the precision with which the weights can be estimated.

Task properties feedback relates to providing forecasters with statistical information on the nature of the task. In time series forecasting, this might involve providing the forecaster with the current estimates of the level, trend and seasonal indices obtained from the Holt–Winters method, for example. However, this would essentially modify the task to one of accepting or adjusting statistical forecasts. Task feedback has been researched widely elsewhere (e.g., Goodwin & Fildes, 1999; Sanders, 1997; Willemain, 1989, 1991), and is not the topic of the current paper.

Ultimately, any form of feedback, regardless of the type, is likely to be most effective in enhancing the judgments of those with technical expertise if it can be understood easily and quickly (O'Connor et al., 2005), and is salient, accurate and timely (Lawrence et al., 2006). We therefore propose and test a rolling training scheme, based on performance feedback. This has a number of innovations that are designed to address the problems associated with feedback that have been presented in earlier studies. Unlike these studies, we have not supplied metrics that summarise the 'average' performance over a given number of periods or tasks (e.g., a mean absolute percentage error or a measure of calibration, which, of necessity, has to be based on a summary of performances over a large number of judgments). Instead, a performance measure is supplied for each individual judgment made by the forecaster, so that there is no arbitrary censoring of earlier performances, and the balance between the sensitivity and stability of the feedback is no longer an issue. Furthermore, the feedback is 'rolling', so that a complete and growing record of the forecaster's performance is presented and updated at regular intervals. These innovations are important because, as we have seen, a key problem with feedback based on 'average' metrics is that it can depend on the number of periods that contribute to the average. Also, when a time series contains cyclical or seasonal patterns, a tendency to forecast too low when the time series rises and too high when it falls will be masked by an 'average' metric. In the scheme proposed here, forecasters can link their errors to individual observations and patterns. They can also see easily whether their performance is improving over time without having to memorise previous values of the metric.

3. Experimental design

3.1. Forecasting approaches

The current research evaluates two judgmental forecasting approaches. Each participant provided judgmental estimates following both approaches, using a fully symmetric experiment, as will be discussed in Section 3.4.

Unaided judgment: This is the simplest judgmental forecasting approach, but is quite popular. Humans are requested to provide point forecasts all at once for all lead times (H), without receiving any kind of guidance, other than the past data points. This approach acts as the benchmark in our study, and is referred to hereafter as UJ.

Rolling training: We propose a direct rolling training approach. Letting N denote the number of observations available for a series and H the number of periods ahead to be estimated, $k > 1$ blocks of H periods each are withheld ($N > kH$). At the first stage, only the first $N - kH$ periods are presented to the forecaster, while H forecasts ahead are requested. When the participants submit their forecasts, the actual values of these H observations are presented, together with performance feedback in terms of percentage errors for each period (signed or not). This procedure is repeated k times, with H data points being added at each repetition. Hence, the completion of each training loop is followed by the submission of H estimates for the future, unknown, periods. As such, we are performing an H-step-ahead rolling evaluation (Tashman, 2000), which is common practice in automatic forecast model selection (Fildes & Petropoulos, 2015). In other words, this is a rolling origin (as opposed to a rolling observation window) forecasting procedure, with updating every H periods, where the observation window is not kept constant but increases with the sample size. In this case, though, instead of selecting the best model based on out-of-sample performances, we assume that this procedure will assist the participants to understand the time series patterns better, thus providing more accurate forecasts. This approach is referred to hereafter as RT.

3.2. Time series

Most relevant studies that have focused on the impact of feedback for judgmental forecasting tasks have made use of simulated series (e.g., Bolger & Önkal-Atay, 2004; Fischer & Harvey, 1999). Moreover, many studies have not examined seasonal series, but have confined their attention to stationary and trended ones (Bolger & Önkal-Atay, 2004; Lurie & Swaminathan, 2009). Therefore, the current research focuses on real time series that collectively demonstrate a variety of characteristics (stationary, only trended, only seasonal, and both trended and seasonal).

More specifically, 16 quarterly series were selected manually from the M3-Competition data set (Makridakis & Hibon, 2000), so as to ensure the required characteristics. These were confirmed using autocorrelation function plots or Cox–Stuart/Friedman tests, or by fitting an appropriate exponential smoothing model, using all the data. In addition, half of the trended or seasonal series did not exhibit any significant pattern (trend or seasonality respectively) in the first two years, but did so later on. This selection was made in order to examine participants' ability to recognise developing series characteristics and adapt.

The 16 series were split into two categories, each containing eight series. These sets of series allowed for the implementation of a symmetric experimental design, which will be described in Section 3.4. Each set contained exactly two series with the same characteristics, as displayed in Table 1. For analysis purposes, the 16 series were split again into two sets of equal size in terms of noise (low and high), as measured by the standardised random component of a classical decomposition. Lastly, four additional series were used in the first (warming-up) stage of the experiment, in order to familiarise the participants with the system.

The required length of all series was set to 28 points (seven years), with longer series being truncated. In both the UJ and RT approaches, the last four observations (last year) were withheld and used only for the out-of-sample evaluation and a comparison of the two approaches. The length of this sample matches the required forecasting horizon; thus, $H = 4$. So, the in-sample consisted of $N = 24$ observations (six years of quarterly data). In addition, 12 observations were used for the RT procedure, thus, the number of blocks was $k = 3$. The forecasting performance was tested on the last four observations (seventh year), with forecasts being produced for both approaches (UJ and RT).

3.3. Participants and web application

The group of participants consisted of 105 undergraduate students who were enrolled in the *Forecasting Techniques* module at the School of Electrical and Computer Engineering at the National Technical University of Athens. As part of the module, the students had been taught principles of time series analysis, statistical and judgmental forecasting methods, and ways of evaluating forecasting performances. The experiment was introduced as an elective exercise, with the 50% of participants who produced the most accurate forecasts obtaining bonus credit.

In order to attract a large number of participants, we decided to build a web application, rather than performing a standard laboratory experiment. The web application was designed specifically for the purpose of this experiment, using the ASP.NET framework for the web development of the front-end and a Microsoft SQL database for storing the time series data and the participants' point forecasts. The Microsoft Chart Controls library was used for drawing line and bar graphs, as is discussed in the next subsection. The application was hosted in a secure web-server and participants could connect remotely through their internet-enabled personal computers via any web browser.

3.4. Process of the experiment

Instead of splitting the participants into two groups, control and test, we adopted a symmetric experimental design, where each participant submitted forecasts for both UJ and RT. The sets of series A and B alternated randomly between UJ and RT, so that half of the series were forecast using UJ by half of the participants and using RT by the other half, and vice-versa for the other series. In order to avoid familiarity with the task, UJ and RT were presented to the participants interchangeably. This means that after a common warm-up round, half of the participants were asked to provide forecasts using the UJ approach for eight time series, then to submit their estimates using the RT approach for the remaining eight series at the next step, while the opposite (first RT then UJ) was the case for the other half of the participants. This symmetric design allowed us to avoid any familiarity with the task effects that could have arisen if the two approaches had been presented in the same order (first UJ then RT) for all of the participants. For the provision of feedback, each participant was assigned randomly to either the signed or unsigned percentage errors treatments (so that either bias or accuracy feedback was provided). Of the 105 participants, 52 were given feedback on signed errors and 53 on absolute errors.

All of the series were presented in a line graph format, using blue for the actual values and green for the submitted forecasts. While there has been no evidence on the relative superiority of graphical or tabular numerical formats (Lawrence et al., 2006), graphical representations are more common in modern forecasting support systems. Historical data points were kept unlabeled in terms of exact values, so that the participants could not export these values to a spreadsheet and use statistical approaches. This is a very important constraint, as the experiment took place in an unobserved environment and a graphical mode of presentation was the only way to guarantee that judgmental extrapolation was used. However, grid lines were provided in order to accommodate numerical estimations. Four text boxes were used for the input of judgmental forecasts for each lead time, while an *update* button could be used to refresh the graph, so that the participant could check his or her judgmental estimates graphically before submitting. Fig. 1 presents two typical screenshots of the system implemented, both before (a) and after (b) the input of the four point forecasts.

Including the warming-up up round, the experiment involved three rounds, each of which is described in detail below. As has been noted, the UJ and RT rounds were presented in reverse order for half of the participants.

Warming-up round: Each of the first four series was presented to the participants in turn, withholding the last four observations. The participants were then requested to provide judgmental point forecasts for the next four quarters (one year). A short description of each series was provided, describing any historical patterns. Once the forecasts for each series had been submitted, forecast errors for each point (signed or not) were calculated automatically and displayed in bar charts, using the color red. As this round was a 'warm-up', the forecasts thus

Table 1
Sets of series.

	Stationary	Trended	Seasonal	Trended and seasonal	Total
Set A	2 series	2 series	2 series	2 series	8 series
Set B	2 series	2 series	2 series	2 series	8 series

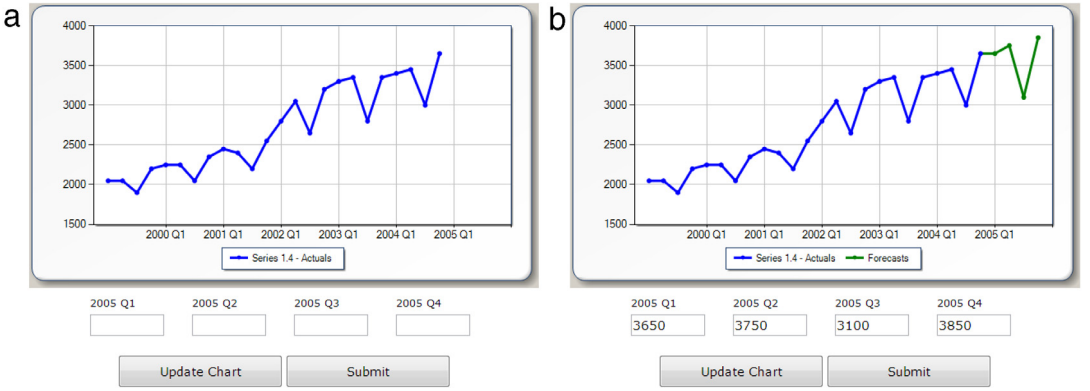

Fig. 1. Screenshots of the system's graphical representation and input features.

elicited were not taken into account when the results of the study were analysed. Fig. 2 presents a screenshot showing the information provided to the participants after the four point forecasts for a series had been submitted.

UJ round: The series from Set A (or Set B) were used, holding back the last four observations in each series. The series were presented in random order. The participants were given the $N = 24$ actual values of each series in a graphical format, and were requested to provide judgmental point forecasts for the next four quarters (periods 25–28). No description of the series or information on the accuracy of the forecasts was provided.

RT round: Series from Set B (or Set A, the opposite of the previous round) were used, holding back the last 16 observations of each series. Again, the series were presented in random order. Each participant was initially given the first $N - kH = 12$ observations and requested to provide four sets of four quarterly judgmental point forecasts, for each of the next four years in a rolling origin manner. First, he or she was asked to submit just the first four point forecasts (for the next year), after which the actual data points were presented, with the corresponding forecast errors (signed or not, as in the warming-up round) being given in a bar chart. Next, the second set of forecasts was requested, followed by the provision of outcome and performance feedback. Then, the third set of forecasts was requested, again followed by outcome and performance feedback. Finally, the participants submitted their last four forecasts. In order to be directly comparable with UJ, only the last set of forecasts was used in the evaluation. Moreover, when producing the forecasts for the final year, participants were given the same amount of information (an observation window of 24 periods) as with the UJ approach.

After completing each of the latter two rounds, the participants filled in a questionnaire, which included questions on their confidence in the accuracy of their submitted forecasts, their expected forecasting performance, the extent to which they had examined the graphs and series patterns, and the time spent in producing their forecasts. In addition, a final questionnaire was used to ask participants about their familiarity with forecasting tasks, their level of forecasting expertise, their perceptions of the effectiveness of rolling training (RT), and their motivation for providing accurate estimates. The two sets of questions posed are given in Table 2. All of the questions had five-step ordinal response choices (Likert scale).

The responses to the questions posed to the participants were analysed in order to discover any relationships between the variables in question (e.g., confidence, expected performance, extent of examination of graphs) and the actual forecasting performances achieved in the respective rounds of the experiment (UJ and RT). The results of this analysis are presented and discussed in Section 4.2.

4. Analysis

4.1. Forecasting performance

Table 3 presents the percentage improvements in accuracy that were achieved by using RT relative to UJ. These percentage improvements are measured as:

$$100 \times \left[1 - median\left(\frac{MAE_s^{\mathrm{RT}}}{MAE_s^{\mathrm{UJ}}}\right)\right] \ (\%),$$

where UJ in the denominator is acting as the benchmark for this study. Negative values indicate that RT performed worse than UJ. In each case, the median is calculated across the series considered. For both the numerator and the denominator, the mean absolute error of a series s is calculated across participants and horizons, as:

$$MAE_s = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{P}\sum_{p=1}^{P}\left|y_h - f_{p,h}\right|,$$

Table 2
Questions posed to the participants.

	Questions
After both UJ and RT rounds	How confident are you that the forecasts you submitted in this round, on average, would be within 10% of the actual values? Please, rate your expected forecasting performance in the series of this round. Did you examine carefully the time series graphs? Did you take into account any historic patterns in the series when making your forecasts during this round? How much time (on average) did you spend for each series of this round? How likely it is that taking more time would change your forecasts?
After completion of the experiment	How familiar are you with such forecasting exercises? How would you describe your level of expertise? Please, rate the effectiveness of rolling training as a tool to increase your accuracy. Please, indicate how motivated you were to provide accurate estimates.

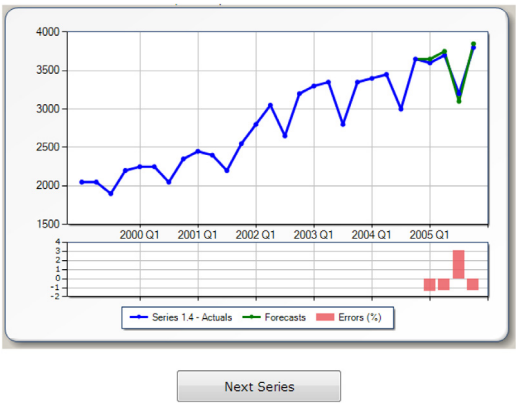

Fig. 2. Screenshot of the system's feedback report features, in terms of outcome (out-of-sample actual values) and performance (error bars).

where P denotes the number of participants, H the number of out-of-sample lead times, y_h the actual value of a series at time h, and $f_{p,h}$ the forecast of participant p for the same series at time h. Note that the number of participants (P) is not the same for all series, due to slight differences in sample sizes.

The results are analysed by columns in terms of series characteristics (stationary, trended, seasonal, trended and seasonal, low noise and high noise). The major rows indicate all (25th–28th), near (25th–26th) and far (27th–28th) horizons. The minor rows provide an additional analysis of the results based on the type of feedback (in the case of RT) provided to the participants. As was mentioned in Section 3.1, two types of feedback have been considered: bias feedback, in the form of signed percentage errors (PE), and accuracy feedback, in the form of absolute percentage errors (APE). Statistically significant differences between RT and UJ have been identified by performing a two-sample paired t-test on the values of the mean absolute errors summarised across participants for each series and each horizon. The analysis was also replicated using the mean absolute percentage error (MAPE) as a measure of the forecasting performance, but no substantial differences in the interpretation of the results were identified.

Overall, there is evidence that the RT approach results in statistically significant better forecasting performances (3.78% performance gain). The improvements are more prominent for high noise (5.18%, statistically significant at the 0.05 level). Although gains of 5.72% and 9.20% were observed for stationary and trended series, respectively, these were not statistically significant at the 0.05 level.

Focusing on the very first row of Table 3, where all horizons are considered, the only case in which RT performs worse comes from the seasonal series. Even though this difference is not statistically significant, suggesting that UJ and RT perform similarly, we attempt to understand the reason behind this result by examining separately series with and without evident seasonality for the very first years, as was discussed in Section 3.2. The results of this analysis suggest that RT might not be suitable for series with developing seasonality.

In terms of the type of feedback provided to the participants, it is apparent that bias feedback demonstrates the most significant improvements (4.89% overall), while the improvements for accuracy feedback are generally smaller and not consistent. One could argue that providing errors in an absolute format may lead to confusion, as the participants may not be able to evaluate this kind of information correctly. On the other hand, bias feedback for each point in the form of signed bar charts is easier to interpret and understand, and indicates a clear strategy for improving one's forecasts. It is notable that bias feedback, which involved the provision of signed percentage errors for each individual period, improved the accuracy for seasonal series. It is unlikely that providing the mean of these percentage errors would have been as effective, because any tendency to over-forecast for some seasons and under-forecast for others would have been masked by the averaging process.

Another very important observation is that RT results in improvements for series both with high noise (5.18%) and when longer horizons are examined (4.17%). These improvement gains are statistically significant at the 0.05 level when all types of feedback are pooled together. However, the differences between RT and UJ are not statistically significant at the 0.05 level for the shorter horizon and low noise series. Lawrence, Edmundson, and O'Connor (1985) suggested that, when the forecasting task is based on graphs, judgmental forecasts can be as good as statistical model forecasts, at least for the shorter horizons. In contrast, unaided judgmental forecasting is likely to be relatively inaccurate for longer horizons and series with high levels of noise. The use of a direct rolling

Table 3
Accuracy improvements (%) of the RT approach over UJ.

	Type of feedback	All series	Stationary	Trended	Seasonal	Trended and seasonal	Low noise	High noise
All horizons (25th–28th)	ALL	3.78*	5.72	9.20	−4.14	0.90	0.90	5.18*
	PE	4.89*	4.10	10.47	2.28	2.58	1.77	5.71*
	APE	3.89	7.27*	7.10	−11.79	0.70	−1.99	6.23
Near horizons (25th–26th)	ALL	2.41	−2.12	−0.91	4.14	8.07*	2.77	2.41
	PE	7.14	0.02	0.63	10.47	7.14	6.50	8.23
	APE	2.04	0.47	−2.45	−3.69	8.71	2.04	0.47
Far horizons (27th–28th)	ALL	4.17*	6.10*	15.74	−10.39	1.59	1.59	6.10*
	PE	5.67	5.67	14.47	−8.14	1.54	6.51	5.67
	APE	2.35	8.14*	12.86	−8.94	−5.47	−2.42	3.50

* The differences are statistically significant at the 0.05 level.

training scheme improves graph-based judgmental long-term forecasting, building on the relative efficiency of judgmental over statistical approaches.

4.2. Questionnaire responses analysis

Fig. 3 provides a graphical representation of the relationships between the participants' responses to the first set of questions (x-axis) and their mean performances (y-axis), as measured by MAPE. Separate lines are presented for UJ (black) and RT (grey). The size of the circle at each data point reflects the number of participants who provided the respective response. As this first set of questions was posed twice (after UJ and RT respectively), we can also examine how the participants alternate their responses after each forecasting approach.

The negative association between the confidence level and MAPE in UJ changes to no correlation for RT. Moreover, participants tend to have fewer expectations for the performances of their submitted forecasts when using RT than UJ. These outcomes are very important, as it is obvious that RT leads participants to be more cautious in their expectations, thus potentially mitigating a well known problem of judgemental forecasting, namely the underestimation of uncertainty (e.g. Makridakis, Hogarth, & Gaba, 2009).

As we would expect, a propensity to examine graphs (and, to a lesser extent, patterns) has a negative association with the MAPE, suggesting that improvements in forecasting accuracy are recorded as participants devote more time to this task. However, literally no differences are observed between the two approaches (UJ and RT) in terms of mean values of the frequency of examining graphs and patterns. One would have expected that RT would motivate the participants to examine the graphs and series patterns more carefully; however, such was not the case.

The forecasting performances achieved with both UJ and RT are associated with the time that the participants reported spending in producing the forecasts for each series—the more time they spent, the greater the accuracy they achieved. However, the correlation is stronger in the case of UJ, meaning that the forecasting performance achieved using the RT approach can be seen as more time invariant. Also, there is evidence that participants who were less accurate recognised that spending more time on the task might have resulted in a change of their forecasts (particularly in the case of the RT group).

The same analysis was performed for the second set of questions. The majority of the participants (76%) found the RT approach to be either effective or very effective. However, familiarity with forecasting exercises, perceived effectiveness of RT and motivation to produce accurate forecasts were associated with the forecasting accuracy only weakly or moderately. Interestingly, participants' self-reported level of expertise had a strong positive association with their realised MAPE, so that those who considered themselves to have greater expertise produced less accurate forecasts. Further work would be needed to establish why this was the case, but it is consistent with the Dunning–Kruger effect (Kruger & Dunning, 1999), where relatively unskilled people mistakenly consider their level of ability to be higher than it really is. Clearly, such an effect would have important implications for EKE if choices are being made between experts' forecasts based on their self-rated expertise.

5. Discussion and implications

The key finding of this study is that, in tasks involving time series extrapolation where no contextual information is available, the judgmental forecasting accuracy of people with a technical knowledge of forecasting can be improved substantially by providing the forecasters with simple, understandable performance feedback. This suggests that training based on feedback can be a valuable element of the EKE process when time series need to be extrapolated. A number of characteristics of this feedback appear to be crucial. First, in order to be most effective, the feedback should relate to bias, rather than accuracy. As was discussed earlier, feedback on bias provides a clear indication of how future forecasts might be improved, whereas feedback on accuracy does not provide any indication of possible improvement strategies. Nor does it provide an indication of whether any improvement in accuracy is even possible. For example, does an APE of 10% represent the limit of the accuracy that can be achieved, given the noise level, or is there scope for further improvement?

Second, the attribute of the bias feedback that appeared to contribute most to its effectiveness was the feedback of a set of individual errors, rather than an average of these errors. In series where the signal has autocorrelated elements, such as seasonal series, judgmental biases may lead to positive errors at some stages of the cycle (e.g., when

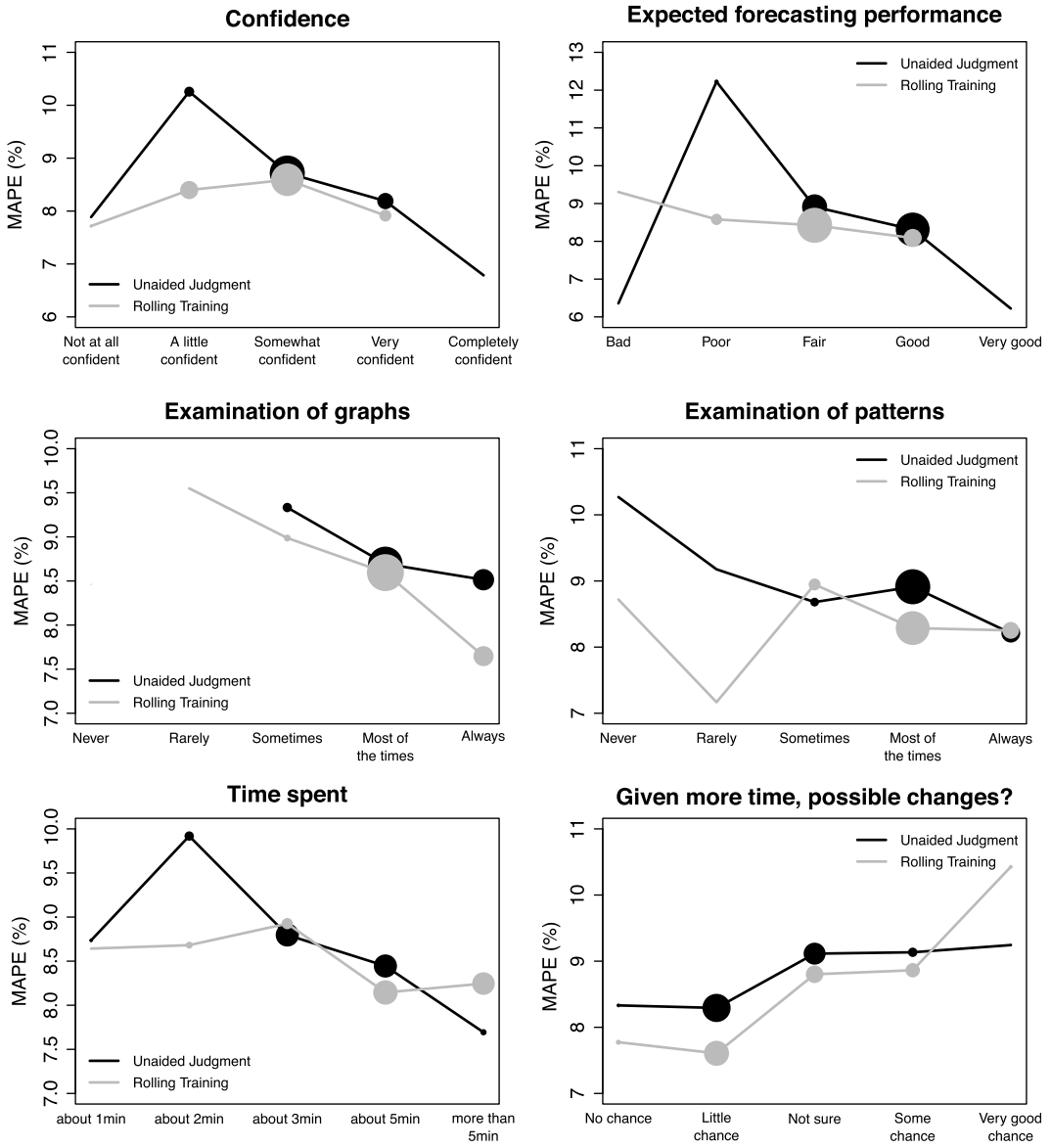

Fig. 3. Association between questionnaire responses and forecasting performances for the first set of questions.

sales are increasing) and negative errors at other stages (e.g., when sales are decreasing). Presenting individual errors allows each observed bias to be associated with a specific period, and avoids the cancelling out of opposing biases that would be a feature of any averaging. Also, the need to select an appropriate length for averaging the point forecast errors is now removed.

Third, presenting the bias feedback as a bar chart may have enhanced its effectiveness, though further research would be needed to establish this. For example, a set of four negative bars would be a strong, simple and clear indication that the previous set of forecasts was too high (*error* = *actual* – *forecast*). A table of four numbers would probably provide a less salient message.

Fourth, the rolling nature of the feedback enabled it to reflect improvements in performance quickly, while at the same time avoiding the danger of confining a participant's attention to the performance of the most recent forecast (which is a danger of outcome feedback). Moreover, rolling across origins for one series before moving on to the second series helped the participants to focus on each series separately and better understand the improvements (or deterioration) in their performance over time. However, this is not a realistic representation of the typical forecasting task; it is more common for feedback to arise across time series.

Recent research suggests that the focus on helping people to learn how to avoid bias is appropriate. A study by Sanders and Graman (2009) found that accuracy was less important than bias when translating forecast errors into costs (such as excessive inventory or labour costs). In their survey of forecasters, Fildes and Goodwin (2007) expressed surprise at the number of company forecasters who never checked the accuracy of their forecasts. The current study

and the findings of Sanders and Graman (2009) suggest that monitoring and feeding back levels of bias may be just as important as checking accuracy levels, or even more so, if the objective is to obtain improved forecasts and minimize the costs of errors.

The proposed RT approach offers an innovative direct feedback approach to time series forecasting. Usually, time series forecasting occurs periodically and across series. Thus, any feedback (lessons learned) from the performance achieved on the previous periods would probably be regarded as outdated. RT offers direct, timely and salient feedback on the performance over a number of periods, focusing on the performance for a single series. Providing the past forecast errors for each period allows specific periods in which the performance dropped to be identified. These two features of RT enable forecasters to achieve better performances for the longer horizons and the more volatile series. This is due to the fact that RT essentially invites the forecasters to examine the patterns in the series closely across a number of horizons, rather than focusing only on short-term forecasts. In addition, as the performance is provided in a rolling manner, forecasters are able to understand the limits of predictability for each series. As such, RT may have an important role to play, being particularly suitable for forecasting and decision making under low levels of predictability (i.e., where there is a high degree of uncertainty).

6. Conclusions and perspectives

Judgmental forecasting is employed in a wide range of contexts for estimating the future values of time series. However, numerous studies have shown the limitations of judgment, even when it is elicited from individuals with technical expertise. The current study has examined the effectiveness of a rolling training scheme that provides direct feedback by reporting to participants their performances for given tasks. This involved reporting signed or absolute percentage errors for each period on a rolling basis, as opposed to metrics that summarise performances over several periods. Real time series featuring a number of characteristics were used. The participants provided estimates for both the control case (unaided judgment) and the test case (rolling training), leading to an increased power. This was achieved using a symmetric experimental design. Although the analysis was not based on data collected in the field, the experimental approach allowed the effects of feedback of different types to be measured and compared efficiently under controlled conditions. Experiments like these have played a valuable role in areas such as behavioural operations management, as one component of a process of triangulation with field research (Siemsen, 2011).

An analysis of the judgmental estimates indicates that a rolling training scheme can improve the accuracy of the judgmental extrapolations elicited from forecasters with technical knowledge, especially when this is combined with feedback in the form of signed errors. Because signed errors indicate the biases in the forecasts, they enable participants' forecasting accuracies to be enhanced. This is particularly obvious in non-stationary series. On the other hand, accuracy feedback based on an absolute form of errors is found to be more difficult to interpret, leading to worse performances in the case of series that exhibit seasonality. Sanders and Ritzman (1992) found little advantage in employing judgmental forecasters with technical knowledge. In contrast, the results presented here suggest that it is worth designing EKE schemes (possibly incorporated into software systems) that build on the technical expertise acquired though didactic learning by providing experiential learning based on feedback that is accurate, timely, suggestive of how improvements might be made, and easy to interpret.

One very interesting finding is that the improvements achieved by using a rolling training procedure are greater for longer forecasting horizons and noisy series. On top of the improvements in forecasting performance achieved, the rolling training procedure also made the participants less confident of their forecasts. This is an additional advantage, as there is evidence that people tend to underestimate the levels of uncertainty associated with their forecasts.

The current paper has focused on analysing performances over the final set of periods (the final year), contrasting unaided judgment with rolling training. However, a further possible objective with the current experimental design would be to analyse how the forecasting performance changes over time within a single series, as a direct result of applying the rolling training procedure. Moreover, policy capturing regression models may provide insights into the forecasting strategies employed by participants with technical knowledge. This could include a large number of potential cues that are linked with time series forecasting. Of course, the time series forecasting task is often carried out in situations where contextual information (such as information from market research or information on advertising strategies) is available to expert forecasters in addition to time series data, and it would be interesting to test the effectiveness of rolling training in this context.

References

Andreassen, P. B. (1988). Explaining the price volume relationship—The difference between price changes and changing prices. *Organizational Behavior and Human Decision Processes*, *41*, 371–389.

Aspinall, W. (2010). A route to more tractable expert advice. *Nature*, *463*(7279), 294–295.

Balzer, W. K., Doherty, M. E., & O'Connor, R. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin*, *106*, 410–433.

Bolger, F., & Harvey, N. (1993). Context-sensitive heuristics in statistical reasoning. *The Quarterly Journal of Experimental Psychology Section A*, *46*, 779–811.

Bolger, F., & Önkal-Atay, D. (2004). The effects of feedback on judgmental interval predictions. *International Journal of Forecasting*, *20*, 29–39.

Bolger, F., & Rowe, G. (2014). Delphi: somewhere between Scylla and Charybdis? *Proceedings of the National Academy of Sciences of the United States of America*, *111*(41), E4284.

Collopy, F. (1994). Review of Nada R. Sanders and Larry P. Ritzman (1992). forecastingprinciples.com reviews of important papers on forecasting [accessed 01.03.15].

Culotta, E. (1992). The calculus of education reform. *Science*, *255*, 1060–1062.

Eggleton, I. R. C. (1982). Intuitive time-series extrapolation. *Journal of Accounting Research*, *20*, 68–102.

Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces*, *37*, 570–576.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, *25*, 3–23.

Fildes, R., & Petropoulos, F. (2015). Simple versus complex selection rules for forecasting many time series. *Journal of Business Research*, *68*, 1692–1701.

Fischer, I., & Harvey, N. (1999). Combining forecasts: what information do judges need to outperform the simple average? *International Journal of Forecasting*, *15*, 227–246.
Franses, P. H. (2014). *Expert adjustments of model forecasts*. Cambridge: Cambridge University Press.
Goodwin, P., & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making*, *12*, 37–53.
Goodwin, P., Onkal-Atay, D., Thomson, M. E., Pollock, A. E., & Macaulay, A. (2004). Feedback-labelling synergies in judgmental stock price forecasting. *Decision Support Systems*, *37*, 175–186.
Goodwin, P., & Wright, G. (1993). Improving judgmental time series forecasting: A review of the guidance provided by research. *International Journal of Forecasting*, *9*, 147–161.
Goodwin, P., & Wright, G. (1994). Heuristics, biases and improvement strategies in judgmental time series forecasting. *Omega International Journal of Management Science*, *22*, 553–568.
Goodwin, P., & Wright, G. (2014). *Decision analysis for management judgment* (5th ed.). Chichester: Wiley.
Harvey, N. (1995). Why are judgments less consistent in less predictable task situations? *Organizational Behavior and Human Decision Processes*, *63*, 247–263.
Klayman, J. (1988). Learning from experience. In B. Brehmer, & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 281–304). Amsterdam: North Holland.
Kremer, M., Moritz, B., & Siemsen, E. (2011). Demand forecasting behavior: System neglect and change detection. *Management Science*, *57*, 1827–1843.
Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*, 1121.
Lawrence, M. J., Edmundson, R. H., & O'Connor, M. J. (1985). An examination of the accuracy of judgmental extrapolation of time series. *International Journal of Forecasting*, *1*, 25–35.
Lawrence, M., Goodwin, P., O'Connor, M., & Onkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, *22*, 493–518.
Lawrence, M. J., & Makridakis, S. (1989). Factors affecting judgmental forecasts and confidence intervals. *Organizational Behavior and Human Decision Processes*, *43*(2), 172–187.
Lawrence, M., & O'Connor, M. (1992). Exploring judgemental forecasting. *International Journal of Forecasting*, *8*, 15–26.
Lawrence, M., & O'Connor, M. (1993). Scale, randomness and the calibration of judgemental confidence intervals. *Organizational Behavior and Human Decision Processes*, *56*, 441–458.
Lopes, L. L., & Oden, G. C. (1987). Distinguishing between random and nonrandom events. *The Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 392–400.
Lurie, N. H., & Swaminathan, J. M. (2009). Is timely information always better? The effect of feedback frequency on decision making. *Organizational Behavior and Human Decision Processes*, *108*, 315–329.
Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, *16*, 451–476.
Makridakis, S., Hogarth, R. M., & Gaba, A. (2009). Forecasting and uncertainty in the economic and business world. *International Journal of Forecasting*, *25*, 794–812.
Moon, J. (2004). *A handbook of reflective and experiential learning: Theory and practice* (p. 126). London: Routledge Falmer.
Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(20), 7176–7184.
Mosteller, F., Siegel, A. F., Trapido, E., & Youtz, C. (1981). Eye fitting straight lines. *The American Statistician*, *35*, 150–152.
Murphy, A. H., & Winkler, R. L. (1977). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? In *National weather digest, Vol. 2* (pp. 2–9).
O'Connor, M., Remus, W., & Griggs, K. (1993). Judgemental forecasting in times of change. *International Journal of Forecasting*, *9*, 163–172.
O'Connor, M., Remus, W., & Griggs, K. (1997). Going up–going down: How good are people at forecasting trends and changes in trends? *Journal of Forecasting*, *16*, 165–176.
O'Connor, M., Remus, W., & Lim, K. (2005). Improving judgmental forecasts with judgmental bootstrapping and task feedback support. *Journal of Behavioral Decision Making*, *18*, 247–260.
Önkal, D., & Muradoglu, G. (1995). Effects of feedback on probabilistic forecasts of stock prices. *International Journal of Forecasting*, *11*, 307–319.
Pollock, A. C., & Wilkie, M. E. (1993). Directional judgemental financial forecasting: trends and random walks. In *Modelling reality and personal modelling* (pp. 253–271). Physica-Verlag HD.
Reimers, S., & Harvey, N. (2011). Sensitivity to autocorrelation in judgmental time series forecasting. *International Journal of Forecasting*, *27*, 1196–1214.
Remus, W., O'Connor, M., & Griggs, K. (1996). Does feedback improve the accuracy of recurrent judgemental forecasts? *Organizational Behavior and Human Decision Processes*, *66*, 22–30.
Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting*, *15*(4), 353–375.
Sanders, N. R. (1997). The impact of task properties feedback on time series judgmental forecasting tasks. *Omega: International Journal of Management Science*, *25*, 135–144.
Sanders, N. R., & Graman, G. A. (2009). Quantifying costs of forecast errors: A case study of the warehouse environment. *Omega: International Journal of Management Science*, *37*, 116–125.
Sanders, N. R., & Ritzman, L. P. (1992). The need for contextual and technical knowledge in judgmental forecasting. *Journal of Behavioral Decision Making*, *5*, 39–52.
Siemsen, E. (2011). The usefulness of behavioral laboratory experiments in supply chain management research. *Journal of Supply Chain Management*, *47*, 17–18.
Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, *16*, 437–450.
Wagenaar, W. A., & Sagaria, S. D. (1975). Misperception of exponential growth. *Perception and Psychophysics*, *18*, 416–422.
Welch, E., Bretschneider, S., & Rohrbaugh, J. (1998). Accuracy of judgmental extrapolation of time series data—Characteristics, causes, and remediation strategies for forecasting. *International Journal of Forecasting*, *14*, 95–110.
Willemain, T. R. (1989). Graphical adjustment of statistical forecasts. *International Journal of Forecasting*, *5*, 179–185.
Willemain, T. R. (1991). The effect of graphical adjustment on forecast accuracy. *International Journal of Forecasting*, *7*, 151–154.

Fotios Petropoulos is Lecturer (Assistant Professor) at Cardiff Business School of Cardiff University. Before that, he was a member of the Lancaster Centre for Forecasting at Lancaster University and the Forecasting and Strategy Unit of the National Technical University of Athens. Fotios is engaged in research on improving forecasting processes.

Paul Goodwin is Emeritus Professor of Management Science at the University of Bath. His research interests are concerned with the integration of management judgment and analytical methods in forecasting and decision making. In 2013 he was elected as an Honorary Fellow of the International Institute of Forecasters. He is co-author of *Decision Analysis for Management Judgment* (Wiley).

Robert Fildes is Distinguished Professor of Management Science in the School of Management, Lancaster University and Director of the Lancaster Centre for Forecasting. He was co-founder of the *Journal of Forecasting* in 1981 and of the *International Journal of Forecasting* in 1985. He has consulted and lectured widely on all aspects of the problem of improving forecasting in organisations.